International Journal of Advanced Engineering Research and Science (IJAERS) *[Vol-3, Issue-12, Dec- 2016]*
https://dx.doi.org/10.22161/ijaers/3.12.21 *ISSN: 2349-6495(P) | 2456-1908(O)*

An Improving Performance of Data Mining Using Intelligent Agent System Methods

Dr. M. Kumarasamy

Professor, Informatics Department, College of Engineering and Technology, Wollega University, Nekemte, Ethiopia.

Abstract— *Our objective of this paper is how we can utilize intelligent techniques in data mining to take decisions which we can use very much in Data Analytics, Credit Card fraud detection and other applications. The author describes that we can develop intelligent software which combines intelligent techniques and data mining and provides solutions for business analysis. Nowadays, Intelligent Agents are widely used in Data Mining because of their flexibility, modularity and general applicability to a wide range of problems. Technological developments in distributed computing, robotics and the emergence of object-orientation have given rise to such technologies to model distributed problem solving. Techniques such as pattern recognition, machine learning, and neural networks have received much attention. Other techniques in Intelligent Agents such as knowledge acquisition, knowledge representation, and search, are relevant to the various process steps in DM. The aim of intelligence is to discover mechanisms of adaptation in a changing environment with utilisation of intelligence, for instance in the ability to exclude unlikely solutions. Intelligence methods have extensive application in different fields such as software, medicine, games and transportation. This paper deals with interdisciplinary issues – interconnection of intelligence and data mining. The main goal of this paper is to point out the intelligence techniques that can be utilised in data mining application and to provide an overview of research undertaken in this field.*

Keywords— Data Mining, Intelligence, Machine Learning.

I. INTRODUCTION

Intelligent Agents, special types of software applications, have become a very popular paradigm in computing in recent years. Some of the reasons for this popularity is their flexibility, modularity and general applicability to a wide range of problems. Recent increase in agent-based applications is also because of the technological developments in distributed computing, robotics and the emergence of object-oriented programming paradigms. Advances in distributed computing technologies has given rise to use of agents that can model distributed problem solving. Besides, object-oriented programming paradigm introduced important concepts into software development process which are used in structuring agent-based approaches. With the explosive growth of information sources available on the Internet, and on the business, government, and scientific databases, it has become increasingly necessary for users to utilize automated and intelligent tools to find the desired information resources, and to track, analyze, summarize, and extract *"knowledge"* from them. These factors have given rise to the necessity of creating server-side and client-side intelligent systems that can effectively mine for knowledge. Therefore, the inherent parallelism and complexity of the classification and discovering patterns from large amounts of data can be delegated to intelligent software agents.

II. INTELLIGENT AGENT

Intelligent Agents are defined as software or hardware entities that perform some set of tasks on behalf of users with some degree of autonomy. In order to work for somebody as an assistant, an agent has to include a certain amount of intelligence, which is the ability to choose among various courses of action, plan, communicate, adapt to changes in the environment, and learn from experience. In general, an intelligent agent can be described as consisting of a *sensing* element that can receive events, a *recognizer* or *classifier* that determines which event occurred, a *set of logic* ranging from hard-coded programs to rule-based inferencing, and a *mechanism* for taking action. Other attributes that are important for agent paradigm include mobility and learning. An agent is *mobile* if it can navigate through a network and perform tasks on remote machines. A *learning* agent adapts to the requirements of its user and automatically changes its behavior in the face of environmental changes. For learning or intelligent agents, an *event-condition-action* paradigm can be defined. In the context of intelligent agents, an *event* is defined as anything that happens to change the environment or anything of

https://dx.doi.org/10.22161/ijaers/3.12.21 *ISSN: 2349-6495(P) | 2456-1908(O)*

which the agent should be aware. For example, an event could be the arrival of a new mail, or it could be a change to a Web page. When an event occurs, the agent has to recognize and evaluate what the event means and then respond to it. This second step, determining what the *condition* or *state* of the world is, could be simple or extremely complex depending on the situation. If mail has arrived, then the event is self-describing, the agent may then have to query the mail system to find out who sent the mail, and what the subject is, or even scan the mail text to find keywords. All of this is part of the recognition component of the cycle. The initial event may wake up the agent, but the agent then has to figure out what the significance of the event in terms of its duties. If the mail is from the boss of the user, then the message can be classified as urgent. This gives the most useful aspect of intelligent agents- *actions*. The main issue in the use of intelligent agents is the concept of autonomy. The user can give the responsibility of performing some time-consuming computer operations to this "smart" software. By this way, the user becomes free to move on other tasks and even disconnect from the computer while the software agent is running. Besides, the user does not have to learn how to do the computer operation. In fact, intelligent agents can act as a layer of software to provide the usability feature that many inexperienced users want from computer professionals.

III. INTEGRATING DATA MINING AND ARTIFICIAL INTELLIGENCE

Data mining is defined as the process of extracting valid, previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions. Data mining is considered as the key process of Knowledge Discovery in Databases (KDD) . The main data mining techniques are Classification and Clustering analysis, Time-series mining, and Association rules mining. Data mining techniques are mostly based on statistics, as well as machine learning while the patterns may be inferred from different types of data. Methods used in data mining, such as machine learning, belong to the field of artificial intelligence. Artificial intelligence (AI) systems are designed to adapt and learn. The first definition of AI is based on the Turing test. Alan Turing undertook a test of a machine‘s ability to demonstrate intelligence. It proceeds as follows: a human judge engages in a natural language conversation with one human and one machine, each of which tries to appear human. The aim of the judge is to distinguish human from machine, only on the basis of conversation (without visual or other help). When the judge cannot distinguish between human and machine, than the machine may be considered as intelligent. The AI approach can be split into two main approaches – Symbolic (Conventional) AI and Sub symbolic AI (Computational intelligence). Conventional AI uses logic and rules to make decisions. Examples of conventional AI techniques are expert systems and Bayesian networks. It is a top-down approach. Computational Intelligence (soft computing) takes inspiration from biological mechanisms and uses a bottom-up approach. Examples of computational intelligence techniques used in economic application are neural networks, genetic algorithms, fuzzy systems etc.

IV. METHODS SHARED IN DATA MINING AND AI

AI research is concerned with the principles and design of rational agents, and data mining systems can be good examples of such rational agents. Most AI research areas have concentrated on the development of symbolic and heuristic methods to solve complex problems efficiently. These methods have also found extensive use in data mining.

- **Symbolic computation**. Many data mining algorithms deal with symbolic values. As a matter of fact, since a large number of data mining algorithms were developed to primarily deal with symbolic values, discretization of continuous attributes has been a popular and important topic in data mining for many years, so that those algorithms can be extended to handle both symbolic and real-valued attributes.
- **Heuristic search**. As in AI, many data mining problems are NP-hard, such as constructing the best decision tree from a given data set, and clustering a given number of data objects into an optimal number of groups. Therefore, heuristic search, divide and conquer, and knowledge acquisition from multiple sources have been common techniques in both data mining and machine learning.

V. CHALLENGES OF DATA MINING

DM is a relatively new field and there are many challenges to be faced. Extracting useful information from data can be a complicated and sometimes a difficult process.

i) Ability to handle different types of data

Many database systems have complex data types, such as hypertext, multimedia data, and spatial data. If a DM technique is robust and powerful, it should be able to perform effective DM on various types of data structures.

https://dx.doi.org/10.22161/ijaers/3.12.21 *ISSN: 2349-6495(P) | 2456-1908(O)*

Though ideal, it is impractical to expect a DM technique to handle all kinds of data and to perform different goals of DM effectively. In general, a specific DM system is built for mining knowledge from a specific kind of data.

ii) Graceful degeneration of DM algorithms

The DM algorithms should be efficient and scaleable. The performance of the algorithm should degenerate gracefully. In other words, the searching, mining, or analyzing time of a DM algorithm should be predictable and acceptable as the size of the database increases.

iii) Valuable DM results

DM system should be able to handle noise and exceptional data efficiently. The discovered information must precisely depict the contents of the database and be beneficial for certain applications. Also, the quality of the discovered information should be interesting and reliable.

iv) Representation of DM requests and results

DM identifies facts or conclusions based on sifting through the data to discover patterns or anomalies. To be effective, the systems should allow users to discover information from their own perspectives and the information should be presented to the users in forms that are comfortable and easy to understand. High level query languages or graphical user interface is required to express the DM requests and the discovered information. End users should be able to specify task commands for the DM system and the results from the DM system should be understandable and usable.

v) Mining at different abstraction levels

It is very difficult to specify exactly what to look for in a database or how to extract useful information from a database. Besides, the value of a piece of information is in the eyes of the beholder ± one person's ``gold mine'' could easily be another person's garbage. To facilitate the mining process, the systems should allow the users to mine at different abstraction levels. For example, a high-level query might disclose an interesting trace that warrants further exploration. Thus, it is important for DM tools to support mining at different levels of granularity.

vi) Mining information from different sources of data

In the ages of the Internet, Intranets, Extranets, and data warehouses, many different sources of data in different formats are available. Mining information from heterogeneous database and new data formats can be challenges in DM. The DM algorithms should be flexible enough to handle data from different sources.

vii) Protection of privacy and data security

DM is a threat to privacy and data security because when data can be viewed from many different angles at different abstraction levels, it threatens the goal of keeping data secured and guarding against the intrusion on privacy. For example, it is relatively easy to compose a profile of an individual (e.g. personality, interests, spending habits, etc.) with data from various sources.

VI. ROLE OF AI IN DATA MINING

In addition to machine learning, other AI fields can potentially contribute significantly to various aspects of the data mining process. We mention a few examples of these areas here:

Natural language presents significant opportunities for mining in free-form text, especially for automated annotation and indexing prior to classification of text corpora. Limited parsing capabilities can help substantially in the task of deciding what an article refers to. Hence, the spectrum from simple natural language

processing all the way to language understanding can help substantially. Also, natural language processing can contribute significantly as an effective interface for stating hints to mining algorithms and visualizing and explaining knowledge derived by a KDD system.

Planning considers a complicated data analysis process. It involves conducting complicated data-access and data-transformation operations; applying preprocessing routines; and, in some cases, paying attention to resource and data-access constraints. Typically, data processing steps are expressed in terms of desired postconditions and preconditions for the application of certain routines, which lends itself easily to representation as a planning problem. In addition, planning ability can play an important role in automated agents to collect data samples or conduct a search to obtain needed data sets.

Intelligent agents can be fired off to collect necessary information from a variety of sources. In addition, information agents can be activated remotely over the network or can trigger on the occurrence of a certain event and start an analysis operation. Finally, agents can help navigate and model the World-Wide Web, another area growing in importance.

Uncertainty in AI includes issues for managing uncertainty, proper inference mechanisms in the presence of uncertainty, and the reasoning about causality, all fundamental to KDD theory and practice.

Knowledge representation includes *ontologies*, new concepts for representing, storing, and accessing knowledge. Also included are schemes for representing knowledge and allowing the use of prior human knowledge about the underlying process by the KDD system. These potential contributions of AI are but a sampling; many

https://dx.doi.org/10.22161/ijaers/3.12.21 *ISSN: 2349-6495(P) | 2456-1908(O)*

others, including human computer interaction, knowledge-acquisition techniques, and the study of mechanisms for reasoning, have the opportunity to contribute to KDD.

VII. APPLYING MACHINE LEARNING TECHNIQUES IN DATA MINING

Machine Learning (ML) Techniques define the ability of a computing machine to improve its performance based on previous results, as arising from the current trends in research publications. Currently its applications to real life problems are extensively developed. Figure 1 presents a list of Machine Learning techniques currently used for data mining problems

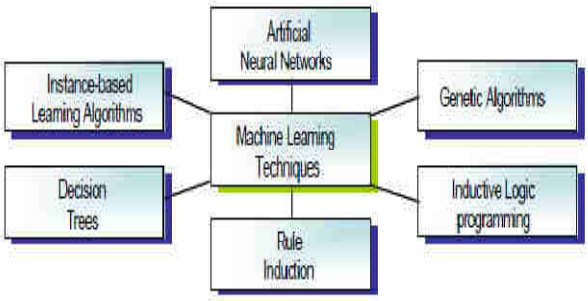

Fig.1: Machine Learning Technique applied to data mining case

From the point of view of applications to DM tasks, ML techniques can be categorized by their usability, performance and impact:

1. **Artificial Neural Networks for DM**: ANN is widely used for DM tasks in many disciplines such as pathology, biology, statistics, image processing, pattern recognition, optimising of numerical analysis as well as controlling systems.
2. **Genetic Algorithms for DM**: Genetic Algorithms, generally evolutionary computation techniques are well-known solving approaches for DM problems in chemistry, biotechnology, movement prediction, bio informatics and adaptive control for working systems.
3. **Inductive Logic Programming for DM**: ILP is shows a restricted area of applications, comparing with other machine learning techniques; however, it has been applied to diagnosis (diseases diagnostic), classification and clustering problems, controlling robotics systems etc.
4. **Rule Induction for DM**: symbolic rule induction shows some applicability for optimisation (a good example would be Semantic query optimisation).
5. **Decision Trees for DM**: DT is a powerful DM tool to solve problems in most of real world cases (prediction, classification etc.). Moreover, by using decision tree induction process, control rules can be derived.
6. **Instance-based Learning Algorithms for DM**: Instance-Based Learning (IBL) is defined as the generalizing of a new instance to be classified from the stored training examples, which is widely used for classification tasks.

VIII. INTELLIGENT AGENT DECISION TECHNOLOGIES

A novel aspect of our work is the use of intelligent decision techniques using data filtering through a supervised or unsupervised feature selection algorithm to select significant features followed by a classifier to improve the performance of data mining. In particular, we investigate the combination of Decision trees, principal component analysis, SPegasos (Stochastic variant of Piramol estimated subgradient solver in SVM), END, Random Forest and Grading for this purpose, which are briefly discussed below.

Decision Trees

In this, the target concept is represented in the form a tree, where the tree is built by using the principle of recursive partitioning. In this, attributes are selected as a partitioning attribute or as a node based on the information gain criteria and then the process continues repeatedly for every child node until all attributes are considered and a decision tree is constructed. Some pruning techniques may further be considered so that the size of the tree is reduced and the overfitting is thereby avoided.

Principal Component Analysis (PCA)

PCA is an unsupervised feature selection based on multivariate statistics and its basic idea is to seek a projection that represents the data in a best possible way in a least-square sense to provide dimensionality reduction. Many researchers pointed out that PCA, which is also known as Karhunen-Loeve transformation in pattern recognition is not found suitable in feature extraction in classification process for the non inclusion of discriminatory information in calculating the optimal rotation of the feature axes.

Stochastic variant of Piramol estimated sub-gradient solver in SVM (SPegasos)

SPegasos implements the stochastic variant of the Pegasos (Primal Estimated sub-GrAdient SOlver for SVM). This implementation globally replaces all missing values and transforms nominal attributes into binary ones. It also

https://dx.doi.org/10.22161/ijaers/3.12.21 *ISSN: 2349-6495(P) | 2456-1908(O)*

normalizes all attributes, so the coefficients in the output are based on the normalized data.

END

END (Ensembles of Balanced Nested Dichotomies for Multi-class Problems) is a Meta classifier for handling multi-class datasets with 2-class classifiers by building an ensemble of nested dichotomies.

Grading

In this type of Meta classifier, the base classifiers are graded to enhance the performance of IDS. We use "graded" predictions (i.e., predictions that have been marked as correct or incorrect) as meta-level classes. For each base classifier, one Meta classifier is learned whose task is to predict when the base classifier results in error. Hence, the way stacking viewed as a generalization of voting, grading may be viewed as a generalization of selection by cross-validation and therefore fills a conceptual gap in the space of meta-classification schemes.

Random Forest (RF)

Random forest is an ensemble of unpruned classification or regression trees, induced from bootstrap samples of the training data, using random feature selection in the tree induction process. Prediction is made by aggregating the predictions of the ensemble by majority voting for classification. It yields generalization error rate and is more robust to noise. However, similar to most classifiers, RF can also suffer from the curse of learning from an extremely imbalanced training data set. As it is constructed to minimize the overall error rate, it will tend to focus more on the prediction accuracy of the majority class, which often results in poor accuracy for the minority class.

IX. ANALYZING AND INTERPRETING DATA

The market environment is continuously changing and demands adequate decision-making from economists - which depends on the application of information technologies. Integrating data and extracting knowledge from the market environment is always complex; in fact, this requires sufficient modeling techniques. maintain that the world "is data rich and information poor", due to the vast amounts of data which are collected but not transformed into information. If this is the case, it could be due to failure to use adequate techniques and technologies to analyze and interpret such data. Thus the data mining implementation phase is critical since there are several techniques that can be considered, especially with the incorporation of AI.

Knowledge representation Data mining seeks to discover interesting patterns from large volumes of data. These patterns can take various forms, such as association rules, classification rules, and decision trees, and therefore knowledge representation becomes an issue of interest in data mining, for instance, trend discovering.

Knowledge acquisition The discovery process shares various algorithms and methods with machine learning for the same purpose of knowledge acquisition from data, or learning from examples, for instance, inductive logic and decision trees.

Knowledge inference The patterns discovered from data need to be verified in various applications and so deduction of mining results is an essential technique in data mining applications", for instance, prediction or forecasting. These three techniques are considered the most common ones associated with AI. This paper established that the most important factor is the integration of artificial intelligence into data warehousing or rather the methodologies that embody AI perspectives.

X. PROPOSED METHODOLOGY

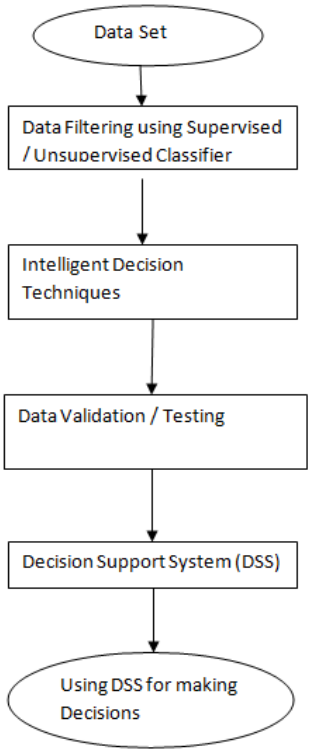

Fig.2: Architecture of Proposed solution

The framework for the proposed architecture is shown in Fig 2 below. Author describes that first of all we have to collect the data sets from various sources. In this, we propose to use combining classifier strategy in order to make intelligent decisions. In this, the data filtering is done

https://dx.doi.org/10.22161/ijaers/3.12.21 *ISSN: 2349-6495(P) | 2456-1908(O)*

after adding supervised classification or unsupervised clustering to the training dataset. Then the filtered data is applied to the final classifier methods to obtain the final decision. After getting the final decision we can keep all the decisions into the decision support system which is very useful to make decisions.

XI. CONCLUSION

The author stated that with using the above proposed software we can predict and solve enormous applications. Knowledge discovery from large volumes of data is a research frontier for both data mining and AI. Knowledge engineering in the 21st century is critical given the huge sums of money invested in business. In business contexts, decision-making requires rich information that can be attained through decision-making support systems. AI has the ability to extract information from the data. Failure to implement such systems can result in organizations making executive decisions based on false alarm prediction, with repercussions that might be detrimental and irreversible. The paper emphasizes the appropriate approach toward designing and implementing AI into data warehousing and data mining. Evidence indicates that there are challenges in relation to knowledge discovery from data, due to lack of application of AI techniques to analyze and interpret the vast amounts of collected data. Notwithstanding, AI techniques that have the ability to address these challenges were highlighted. In fact, it was put forward that these techniques are applicable as per domain since each has been developed to address problems of a specific domain: for example, financial domain, health domain and agricultural domain. In this paper, we investigated some intelligent decision technologies using data filtering by adding supervised or un-supervised methods along with a classifier to make intelligent decisions to support the business organizations.

REFERENCES

[1] G.Guo, D.Neagu and M.T.D.Cronin. A Study on Feature Selection for Toxicity Prediction. Proceedings of FSKD'05 (in press), 27-29 August, Changsha, China, 2005.

[2] Wang, S. L., Patel, D., Jafari, A. & Hong, T. P.(2007). Hiding collaborative recommendation association rules. Applied Intelligence, vol. 26, no. 1,pp. 66-77.

[3] Agrawal, R., Imielinski, T. & Swami, A. (2013). Mining association rules between sets of items in large databases. ACM SIGMOD Conference on Management of Data, pp. 207–216.

[4] Zeng, Y., Yin, S., Liu, J. & Zhang, M. (2015). Research of Improved FP-Growth Algorithm in Association Rules Mining. Scientific Programming, http://dx.doi.org/10.1155/2015/910281.

[5] Hai, L. Q., Somjit, A. & Ngamnij, A. (2013). Association rule hiding based on intersection lattice. Mathematical Problems in Engineering, http://dx.doi.org/10.1155/2013/210405.

[6] Hai, L. Q. & Somjit, A. (2012). A conceptual framework for privacy preserving of association rule mining in e-commerce. 7th IEEE Conference on Industrial Electronics and Applications, pp. 1999–2003.

[7] Hai, L. Q., Somjit, A., Huy, X. N. & Ngamnij, A. (2013). Association rule hiding in risk management for retail supply chain collaboration. Computers in Industry, vol. 64, pp. 776–784.

[8] Truta, T. M. & Campan, A. (2010). Avoiding Attribute Disclosure with the (Extended) p-Sensitive k-Anonymity Model, In: Stahlbock, R., and Crone, S. F.and Lessmann, S. (Eds.), Data Mining: Special Issue in Annals of Information Systems, pp. 353-373.

[9] Jena, L. K., Kamila, N. & Mishra, S. (2014). Privacy preserving distributed data mining with evolutionary computing. Advances in Intelligent Systems and Computing, vol. 247, pp. 259-267.

[10] Ogor, E. N. (2007). Student academic performance monitoring and evaluation using data mining techniques. Electronics, Robotics and Automotive Mechanics Conference (CERMA 2007), pp. 354-359.

[11] Panda M and Patra MR. Bayesian, "Belief network with genetic local search for detecting network intrusions", International journal of secure digital information age 2009; 1(1):34-44.

[12] Tavallaee M, Stakhanova N and Ghorbani AA., "Towards credible evaluation of anomaly based intrusion detection methods", IEEE Transaction on System, Man and Cybernetics, Part-c, Applications and Reviews, 2010; 40(5):516-24.

[13] Chan TS, Yen KK and Luo J., "Network intrusion detection design using feature selection of soft computing paradigms", International journal of computational intelligence ,2008, 4(3):196-208.

[14] R. J. Urbanowicz and J. H. Moore. Learning Classifier Systems: A Complete Introduction, Review, and Roadmap. Journal of Artificial Evolution and Applications, 2009(1):1–25, 2009.

https://dx.doi.org/10.22161/ijaers/3.12.21 *ISSN: 2349-6495(P) | 2456-1908(O)*

[15] S. W. Wilson. Learning classifier systems. Chapter Classifier Conditions Using Gene Expression Programming, pages 206–217. Springer, 2008.
[16] Mansoori, E. G. (2011). FRBC:A fuzzy rule-based clustering algorithm. IEEE Transactions on Fuzzy Systems, Vol. 19, No. 5, pp. 960-971.
[17] O. Parsons and G. A. Carpenter. ARTMAP neural network for information fusion and data mining: map production and target recognition methodologies. Neural Networks, 16:1075–1089, 2003.
[18] A. Davidson J. Schaeffer, D. Billings and D. Szafron. The challenge of poker. Artificial Intelligence Journal, 134(1-2):201–240, 2002.
[19] Stuart Russell and Peter Norvig. Artificial Intelligence: A Modern Approach.Pearson Education, second edition, 2003.
[20] Frank, A. & Asuncion, A. (2010). UCI machine learning repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California', School of Information and Computer Science.
[21] Kulesza, T., Stumpf, S., Wong, W.-K., Burnett, M., Perona, S., Ko, A., Oberst, I. Why-oriented end-user debugging of Naïve Bayes text classification. Transactions on Interactive Intelligent Systems 1(1), ACM (2011).